

NeurIPS 2026 Spotlight

Revisiting Embodied Chain-of-Thought for Generalizable Robot Manipulation

Reasoning supervision for direct robot control

Nan Sun^{1,2}, Yuan Zhang^{2,†,*}, Yongkun Yang³, Wentao Zhao¹, Peiyan Li^{2,4},
Jun Guo^{1,2}, Wenxuan Song^{2,5}, Pengxiang Ding^{2,6}, Runze Suo^{2,7,9},
Yifei Su², Xin Xiao⁸, Xinghang Li², Huaping Liu^{1,*}

¹Tsinghua University ²Xiaomi Robotics ³Peking University ⁴CASIA ⁵HKUST(GZ)

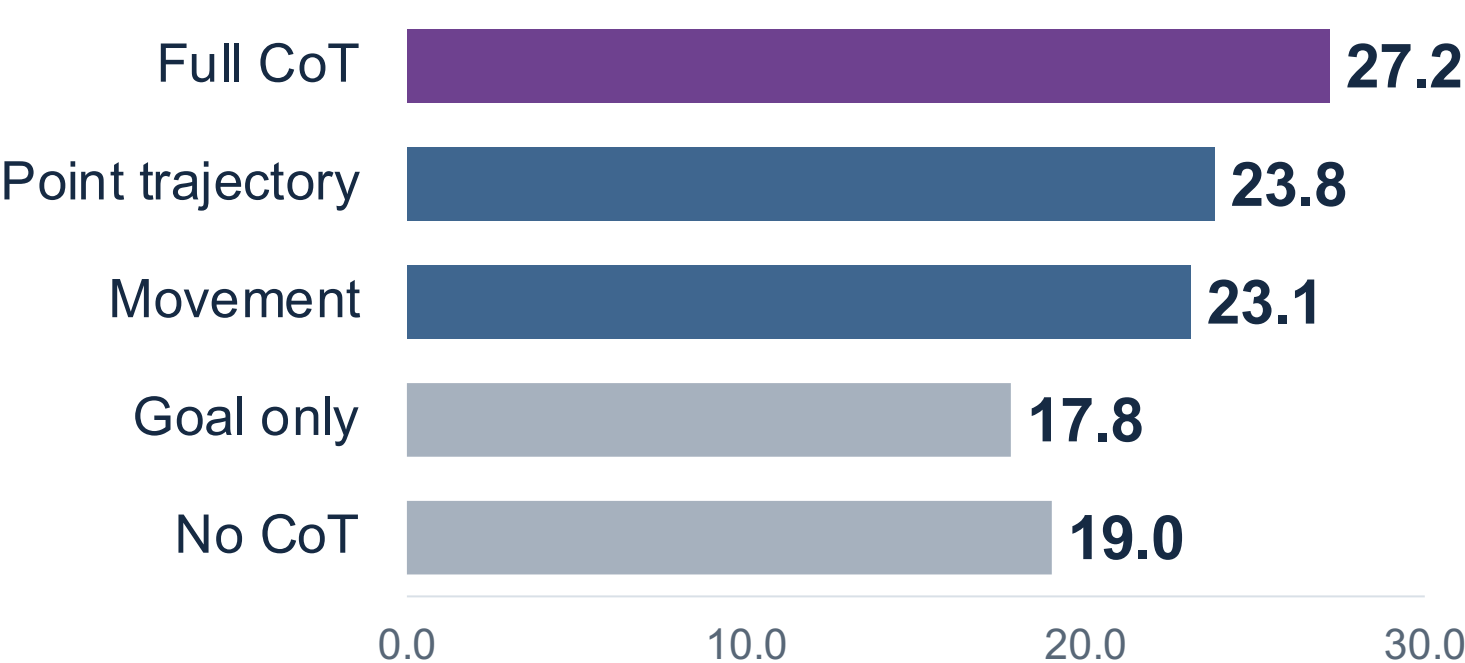
⁶Zhejiang University ⁷Fudan University ⁸Wuhan University ⁹Shanghai Innovation Institute

†Project Leader *Corresponding Author

Action-grounded CoT improves robot control

Which reasoning signals help the policy?

VLABench success rate (%)



+8.2

percentage points

Full CoT over no CoT

Movement and 2D paths connect intent to motion.

Controlled study: Qwen3-VL-4B with a fixed CoT/FAST decoder, without robot pre-training. These values describe the field study, not the final ERVLA model.

A corpus that links semantics to action

978,743

trajectories

226.3M

samples

2,592.5

robot hours

Embodied CoT Format(One Step)

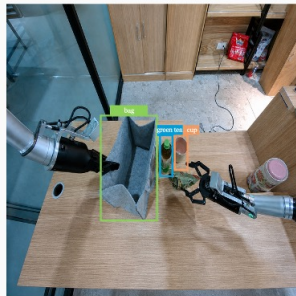
Understanding

goal:

Put the green tee which is on the right side of the table into the bag on the left

Grounding

visible_objects:



Green tee: (x1, y1, x2, y2)



Planning

plan:

1. Approach the drink
2. Pick up the drink
3. Move onto the bag on the left
4. Place the drink into the bag

subtask:

Approach the drink
reasoning:
The drink is on the right of the table, I first need to pick up the drink bottle



Acting

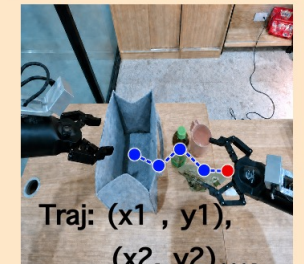
movement_description:

move forward 3cm, then close gripper

gripper_pose:



point trajectory:



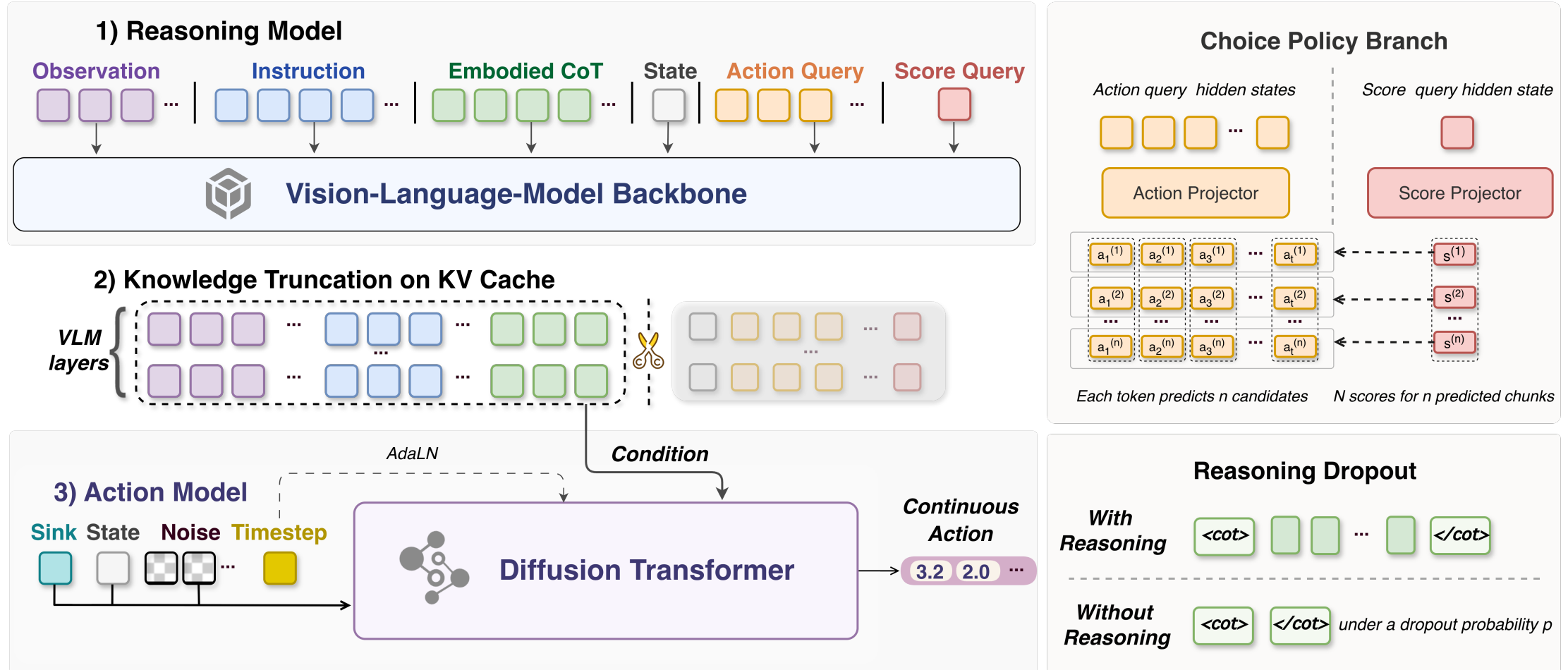
Five robot datasets. Multiple views. Single-arm and bimanual manipulation.

ERVLA learns reasoning in action representations

Choice supervision

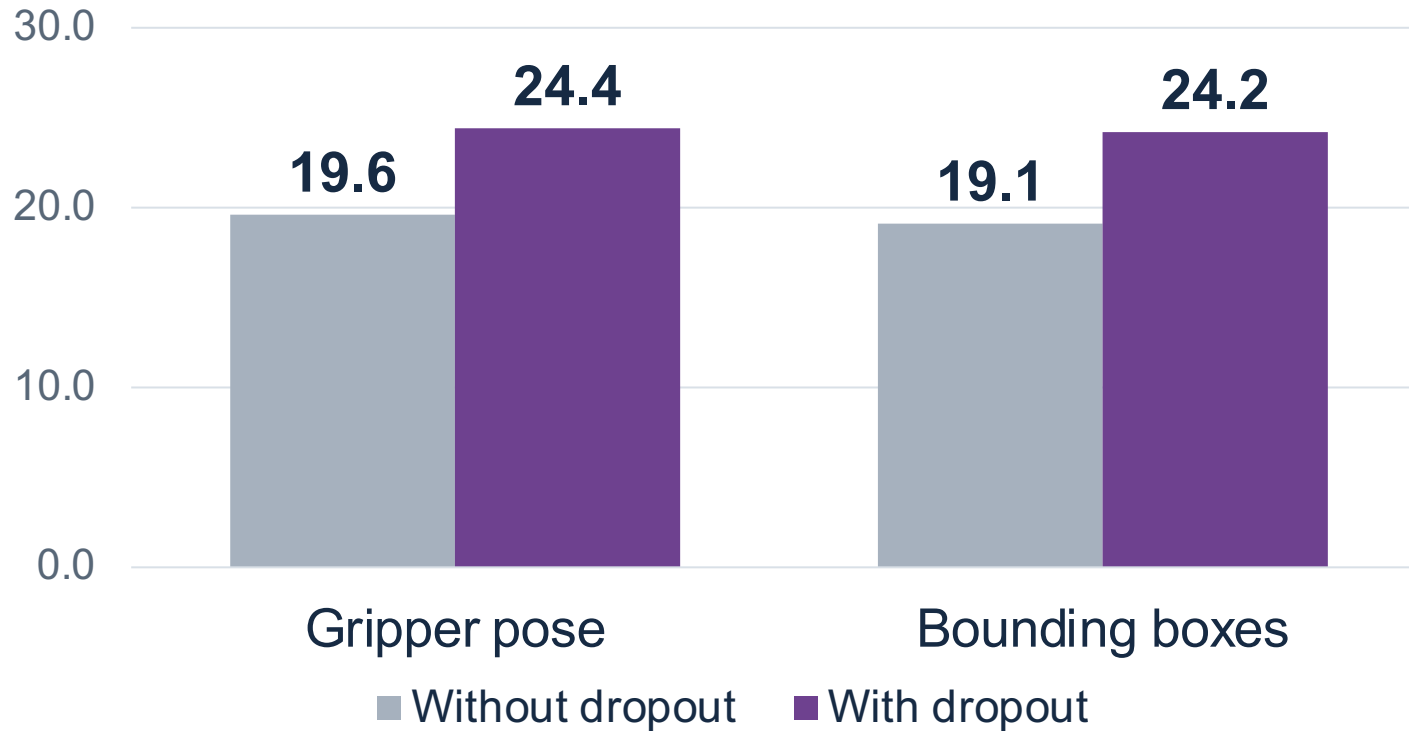
Semantic KV memory

Joint action gradients



Reasoning dropout limits damage from noisy labels

VLABench success rate (%)



$p = 0.5$

CoT dropout in ERVLA training

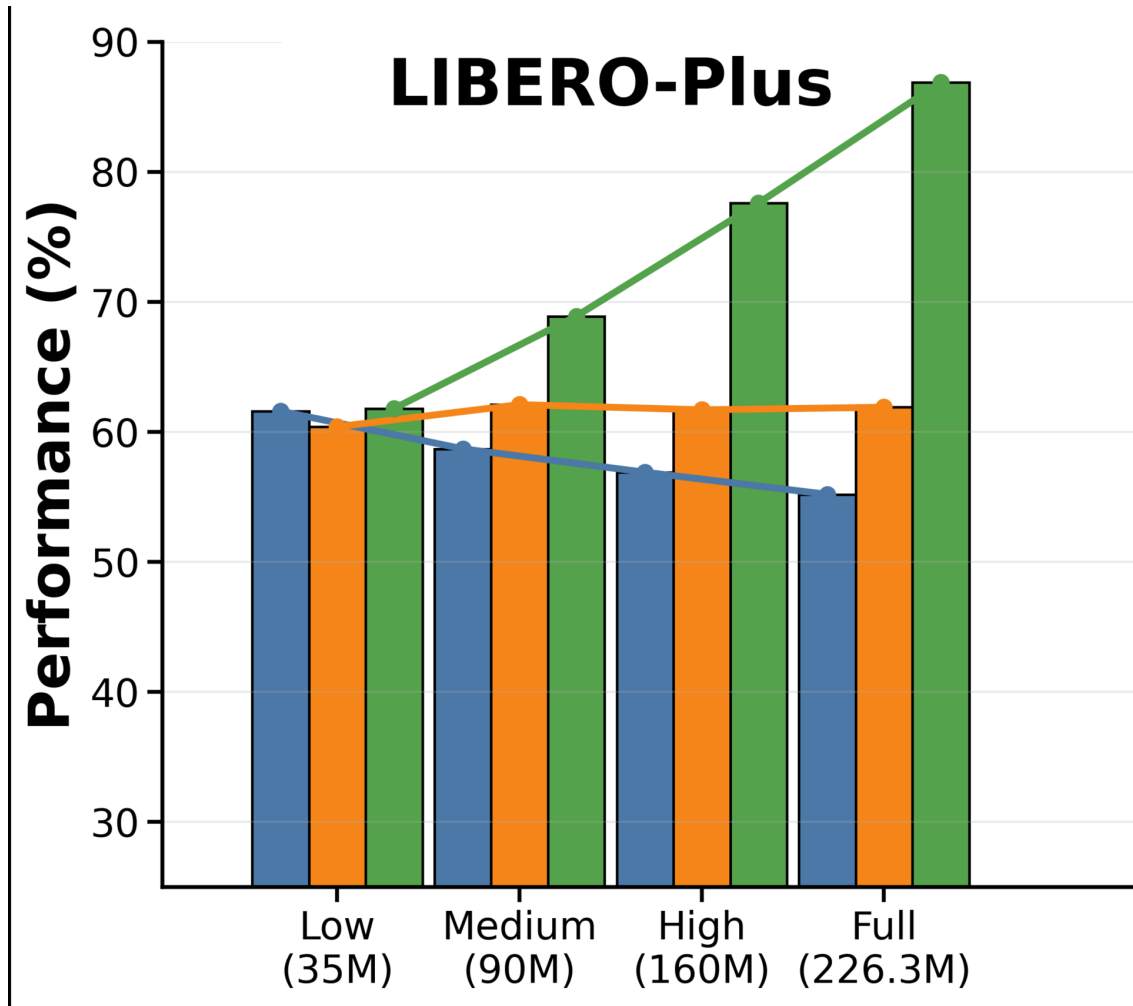
Direct actions at inference

Explicit CoT remains optional

No-CoT reference: 25.2%

Controlled CoT/FAST study with Bridge pre-training. Reasoning dropout mitigates noisy grounding. ERVLA supports direct actions and optional explicit or sparsely refreshed CoT.

ERVLA improves as embodied CoT data grows



35M to 226.3M

embodied CoT samples

ERVLA + Choice

Steadily improves

AR CoT + FAST

Performance falls with more data

VLM + DiT

Largely plateaus

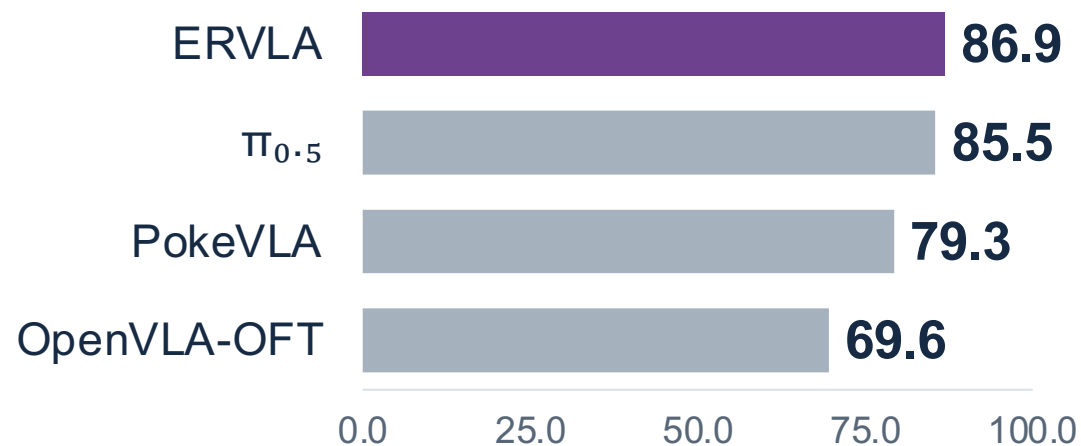
LIBERO-Plus scaling study, original paper Figure 5. The plot preserves the reported performance axis.

Generalization across simulation benchmarks

86.9%

LIBERO-Plus

Total success rate (%)



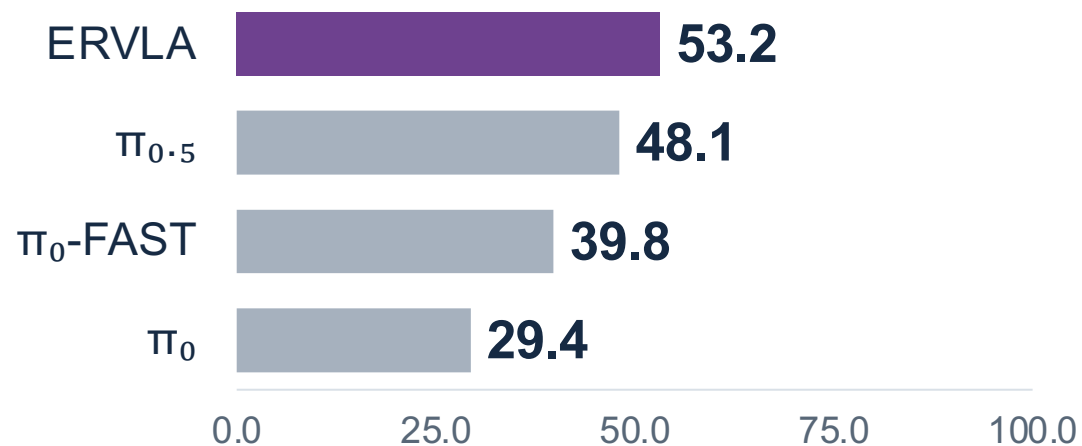
+1.4 pp above $\pi_{0.5}$

LIBERO-Plus: zero-shot transfer after LIBERO post-training. VLABench: mean over five tracks.

53.2%

VLABench

Average success rate (%)



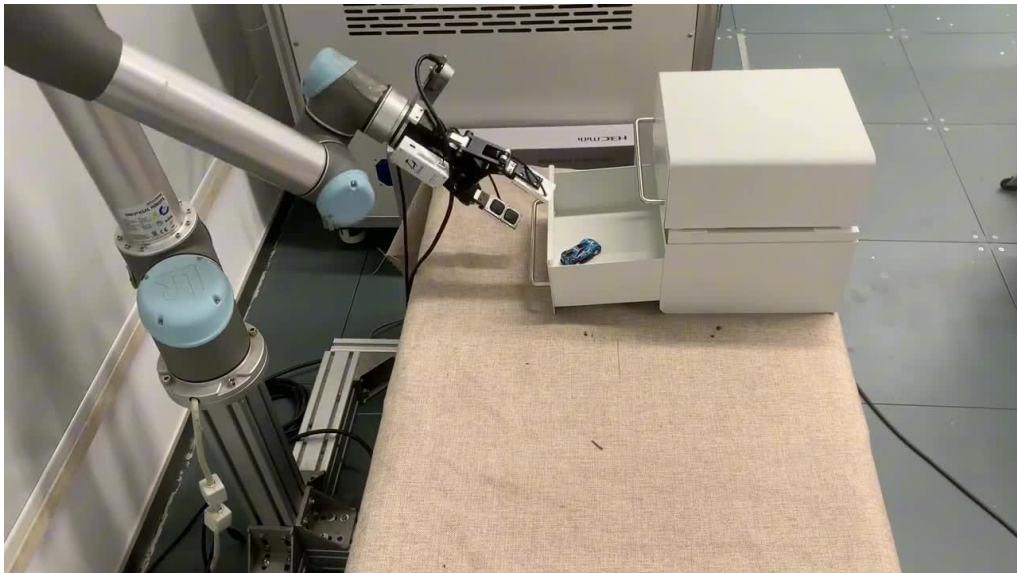
+5.1 pp above $\pi_{0.5}$

Semantic and long-horizon gains on real robots

20 tasks, 100 rollouts/method. Tier means across task families. Progress score measures partial completion.

Semantic instruction

[Click to play](#)



Lower drawer instruction. 2× speed.

Semantic success rate

31%

$\pi_{0.5}$

42%

ERVLA

Long-horizon sorting

[Click to play](#)



Toys below, non-toys above. 3× speed.

Long-horizon progress score

38

$\pi_{0.5}$

55

ERVLA

LIBERO-Plus: a shifted camera viewpoint



Click to play 3× speed

**Pick up the book and place it
in the back compartment
of the caddy.**

77.2%

Camera perturbation success

Zero-shot LIBERO-Plus transfer
after LIBERO post-training

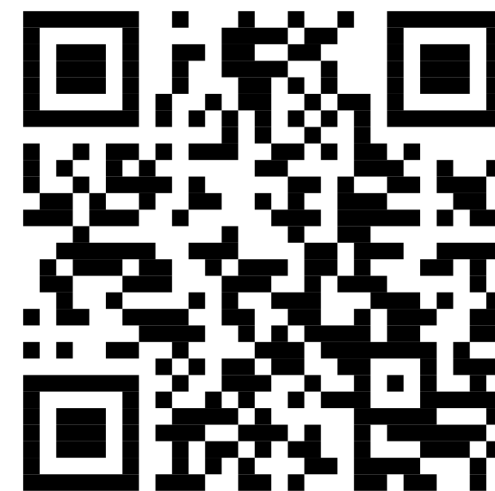
Embodied CoT shapes action representations

Direct actions at inference.
Explicit CoT stays optional.

Action-grounded supervision supports transfer.
Joint action feedback supports data scaling.

Annotation quality and adaptive reasoning
remain open directions.

[arXiv:2606.03784](https://arxiv.org/abs/2606.03784)



taoshuaiz.github.io/ERVLA/